基于直接法的视觉同时定位与地图构建技术综述 *

潘林豪 [1]，田福庆 [1]，应文健 [1]，邱千钧 [2]

(1. 海军工程大学 兵器院，武汉 430033; 2. 海军驻航天二院代表室，北京 100854)

摘　要：视觉同时定位与地图构建（V-SLAM）在机器人、无人机导航、自动驾驶等领域有着广泛的研究。直接法 V-SLAM 基于环境亮度不变性假设，跟踪相机的位姿并构建环境地图。针对直接法 V-SLAM，首先简述其基本原理；然后分析、比较几种具有代表性的直接法 V-SLAM 系统；最后讨论直接法的优缺点和发展趋势，并进行了总结和展望。

关键词：计算机视觉；同时定位与地图构建；直接法；运动推断结构；多视图几何

中图分类号：TP242.62　　doi: 10.3969/j.issn.1001-3695.2018.01.0123

Survey on direct-method visual simultaneous localization and mapping

Pan Linhao[1], Tian Fuqing[1], Ying Wenjian[1], Qiu Qianjun[2]

(1. *College of Weapon*, *Naval Engineering University*, *Wuhan* 430033, *China*; 2. *Naval Representative Office in the* 2nd *Academy of CASA*, *Beijing* 100854, *China*)

Abstract: Visual simultaneous localization and mapping has a wide range of research in the fields of robot, unmanned aerial vehicle navigation, automatic driving, et al. Direct-method V-SLAM, based on the assuption of photometric consistency, tracks the camera’s position and orientation and builds an environment map. This paper, for the direct-method V-SLAM, summarizes the basic principles , and introduces some state-of-art direct-method V-SLAM systems with analysis and comparison. Finally, it discusses the advantages and disadvantages of direct-method V-SLAM and some research tendency, as well as making conclusion.

Key words: computer vision; simultaneous localization and mapping; direct-method; structure-from-motion; multiple-view geometry

0　引言

同时定位与地图构建（simultaneous location and mapping，SLAM）是指搭载传感器的主体，在没有环境先验信息的情况下，于运动过程中跟踪主体的位置与姿态（下文简称“位姿”），并同时构建环境结构一致性地图的技术[1-2]。而视觉同时定位与地图构建（V-SLAM）所用的传感器主要为相机[3]。近年来视觉同时定位与地图构建技术已经在室内服务机器人、自动驾驶汽车、无人机导航定位以及增强现实设备[4-6]中得到了一定的应用。例如，智能手机利用摄像头和 imu 实时定位设备在环境中的位姿，实现 AR 效果；无人机利用摄像头跟踪自身位姿、构建环境地图实现自主飞行。

从 1986 年提出至今，同时定位与地图构建问题已经有超过三十年的研究。该技术源自于机器人领域，在最早涉及 SLAM 的文章中，研究人员把它称为空间状态的不确定性估计（estimation of spatial uncertainty）[7,8]。在早期的研究中，SLAM 问题被当作一个状态估计问题。在这一时期，滤波算法在 SLAM 研究中占据主导地位。研究人员应用状态估计理论，表示传感器位姿和环境结构的不确定性，然后应用滤波器更新优化状态估计的均值和方差[9]。为解决状态方程的非线性问题，研究人员将扩展卡尔曼滤波（EKF）、粒子滤波（PF）、无迹卡尔曼滤波（UKF）等方法应用于 SLAM 问题中。该时期的代表性成果有 EKF-SLAM、UKF-SLAM、FastSLAM 等[10~12]。进入 21 世纪，相机开始在 SLAM 研究中得到广泛应用。Davison 在 2003 年提出了 MonoSLAM[13]，该系统基于单目相机和 EKF 框架最早实现了实时运行的 V-SLAM 系统。同时在该时期，研究人员发现计算机视觉领域的运动推断结构（structure-from-motion，SFM）问题与 SLAM 问题有许多共同点[14]。SFM 技术中的非线性优化方法——捆集调整（bundle adjustment，BA）[15]被引入 SLAM 研究中，成为 V-SLAM 中的主导方法。基于关键帧与捆集调整的 PTAM[16-17]是第一个以非线性优化作为后端的 V-SLAM 系统。它开创性地将相机位姿跟踪（tracking）与地图构

收稿日期：2018-01-21；**修回日期：**2018-03-15　　**基金项目：**海军工程大学自主立项基金资助项目

作者简介：潘林豪（1991-），男（通信作者），浙江温州人，博士研究生，主要研究方向为计算机视觉、深度学习（jaypancool@gmail.com）；田福庆（1962-），男，教授，博士，主要研究方向为自动控制原理；应文健（1979-），男，讲师，博士，主要研究方向为计算机视觉、机器人控制；邱千钧（1990-），男，助理研究员，硕士，主要研究方向为自动控制原理.a http: /rpg.ifi.uzh.ch/.

建（mapping）作为两个线程并行处理，减小相机位姿跟踪的计算量、优化地图构建的精度，成为V-SLAM中里程碑式的工作。之后提出的 V-SLAM 系统，无论是基于特征点法（如 ORB-SLAM[18-19]）还是基于直接法（如 LSD-SLAM[20]）都借鉴了 PTAM 的算法框架。

经过三十多年的研究，V-SLAM 技术形成了以视觉里程计为前端，非线性优化与滤波为后端，搭配回环检测和建图模块的稳定框架。特征点法和直接法是V-SLAM中的两种主流方法。与特征点法相比，直接法的研究时间较短。但经过近几年的发展，直接法在 SLAM 问题的研究中取得了突破，实现了与特征点法相近或更好的效果。本文专注直接法 V-SLAM，系统地分析和比较目前几种代表性的直接法 V-SLAM 系统，讨论直接法的优缺点和发展趋势，并做总结和展望。

1 V-SLAM 基本原理

V-SLAM 技术可以通过相机拍摄的图像序列实时跟踪相机的位姿，并构建环境三维地图。为了跟踪相机的运动，以变换矩阵 $\boldsymbol{T}$ 表示相机的位姿[14]，场景的三维结构由空间点 $\boldsymbol{X}\left(\boldsymbol{X}\in\mathbb{R}^3\right)$ 表示。通过观测方程

$$\boldsymbol{Z}_{ij}=h_{ij}\left(\boldsymbol{T}_i,\boldsymbol{X}_j\right)+\boldsymbol{n}_{ij} \tag{1}$$

可以得到空间点 $\boldsymbol{X}_j$ 在相机 $\boldsymbol{T}_i$ 图像上的观测值 $\boldsymbol{Z}_{ij}$，$\boldsymbol{n}_{ij}$ 表示观测噪声。为了得到空间点与相机的位姿，应用极大似然估计得到状态估计值 $\hat{x}$

$$\hat{x}=\arg\max_x p(z|x)=\arg\max_{\boldsymbol{T}_i,\boldsymbol{X}_j}\prod p\left(\boldsymbol{Z}_{ij}\middle|\boldsymbol{T}_i,\boldsymbol{X}_j\right) \tag{2}$$

由于观测噪声的存在，上述问题可以转换为求解如下目标函数：

$$\arg\max_{\boldsymbol{T},\boldsymbol{X}}\sum_{i=1}^{m}\sum_{j=1}^{n}\left\|\boldsymbol{Z}_{ij}-h_{ij}\left(\boldsymbol{T}_i,\boldsymbol{X}_j\right)\right\|_{\boldsymbol{\Sigma}_{ij}}^2 \tag{3}$$

如图 1 所示，特征点法和直接法用不同的方式参数化公式（3）中的残差。特征点法通过在图像中提取特征点，并匹配特征点周围的描述子，得到对应的特征点 $\boldsymbol{h}_{11}$ 和 $\boldsymbol{h}_{21}$ 之间的重投影误差 $\boldsymbol{e}$。直接法不进行特征点和描述子的匹配，而是以相机位姿估计值为初始值，依据像素梯度寻找与像素点 $\boldsymbol{h}_{11}$ 对应的像素点 $\boldsymbol{h}_{21}$ 的位置。通过优化光度误差 $I(\boldsymbol{h}_{11})-I(\boldsymbol{h}_{21})$ 求解最优的相机位姿[21]。

2 直接法 V-SLAM 系统

目前，直接法 V-SLAM 系统主要有以普通相机为传感器的 SVO(Semi-direct Visual Odometry)[22]、DSO(direct sparse odometry)[23]、LSD-SLAM(large scale direct SLAM)[20]、DTAM(Dense Tracking and Mapping)[24],以及以 RGB-D 相机为传感器的 DVO[25]。本节分析几种具有代表性的单目直接法 V-SLAM 系统，并比较其优缺点。

2.1 SVO 算法分析

SVO 是 Forster 等人于 2014 年提出的一种稀疏直接法系统。在 2017 年扩展了大视角相机、多相机功能，并融合了 imu 惯性器件[26]。SVO 最显著的特点是，前端基于环境亮度不变性假设跟踪相机位姿，后端仍使用传统的最小化重投影误差的方法优化全局地图，因此它也被称做半直接法，重构效果如图 2 所示。

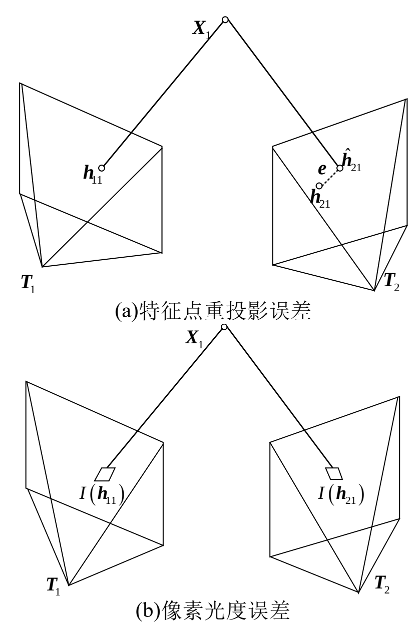


图 1 特征点法与直接法的残差

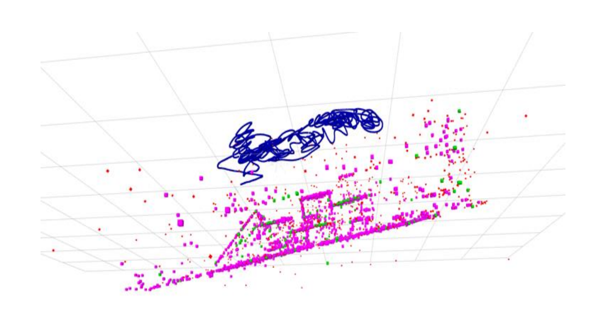


图 2 SVO 重构地图

单目 SVO 算法分为相机位姿跟踪与地图构建两个线程，如图 3 所示。在相机位姿跟踪线程中，SVO 将当前帧 k 的位姿近似为上一帧 $k-1$ 的位姿。将 $k-1$ 帧保存的稀疏特征点投影到当前帧 k 中，通过最小化前后两帧特征点的光度误差，得到两帧之间的相对运动估计值 $\boldsymbol{T}_{k,k-1}$

$$\boldsymbol{T}_{k,k-1}=\arg\min_{\boldsymbol{T}_{k,k-1}}\frac{1}{2}\sum_{i\in R}\left\|\delta I\left(\boldsymbol{T}_{k,k-1},\boldsymbol{u}_i\right)\right\|^2 \tag{4}$$

其中 $\delta I(\cdot)$ 为光度误差，

$$\delta I\left(\boldsymbol{T}_{k,k-1},\boldsymbol{u}_i\right)=I_k\left(\omega\left(\boldsymbol{u}_i,d_{k-1}(\boldsymbol{u}_i),\boldsymbol{T}_{k,k-1}\right)\right)-I_{k-1}(\boldsymbol{u}_i) \tag{5}$$

为了消除特征点与位姿估计的累计误差，将存储在关键帧集合 $r_{1,\cdots,n}$ 中的特征点投影到当前帧 k 中，再通过最小化投影块光度误差，得到特征点在当前帧中的优化投影位置：

$$\boldsymbol{u}_i'=\arg\min_{\boldsymbol{u}_i'}\frac{1}{2}\left\|I_k(\boldsymbol{u}_i')-\boldsymbol{A}_i\cdot I_r(\boldsymbol{u}_i)\right\|^2,\quad\forall i \tag{6}$$

最后，最小化特征点位置 u_i' 与预测的特征点位置之间的重投影误差，分别得到优化的相机位姿和特征点空间位置。

$$T_{kw}=\arg\min_{T_{kw}}\frac{1}{2}\sum_i\left\|u_i'-\omega(T_{kw},{}_w p_i)\right\|^2 \quad (7)$$

其中：T_{kw} 为当前时刻相机在世界坐标系中的位姿；${}_w p_i$ 为特征点在世界坐标系中的空间位置。

为构建场景地图，SVO 的后台地图线程持续恢复参考帧 r 中特征点的空间位置。当有效特征点集合 $\bar{R}$ 的数量小于规定阈值，选取新关键帧作为参考帧 r 并提取特征点。在随后观测到的各帧中利用极线搜索匹配特征点的位置，并通过三角化得到特征点的逆深度[27]观测值 $\hat{\rho}^k$。在 SVO 中，利用一种高斯分布加均匀分布"的模型表示逆深度观测值的分布[28,29]，

$$p\left(\hat{\rho}^k\mid\rho,\gamma\right)=\gamma N\left(\hat{\rho}^k\mid\rho,\tau^2\right)+(1-\gamma)U\left(\hat{\rho}^k\mid\rho_{\min},\rho_{\max}\right) \quad (8)$$

其中；γ 为特征点是内点的概率，τ^2 为内点观测的方差，ρ 为逆深度估计值；$[\rho_{\min},\rho_{\max}]$ 为外点分布的范围。SVO 融合各帧中特征点的测量信息，当方差小于设定阈值，就根据估计的深度值将特征点保存至环境结构地图中，供跟踪线程使用。

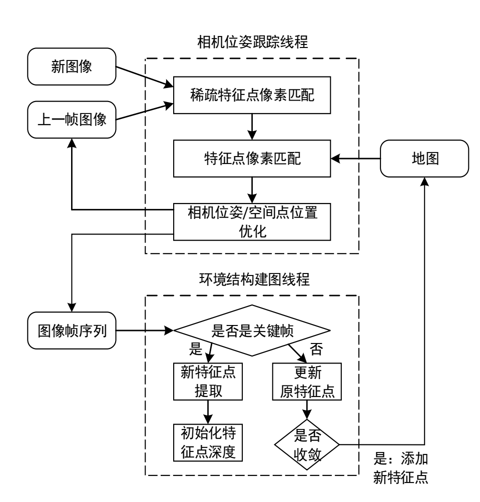
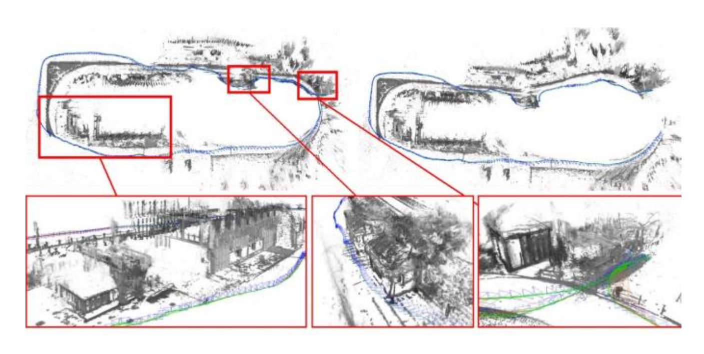


图 3　SVO 算法框架

2.2　LSD-SLAM 算法分析

LSD-SLAM 是一种比较完备的 SLAM 系统。基于前期单目半稠密视觉里程计的工作[30]，Engel 在 2014 年提出了单目 LSD-SLAM 系统[20]。此后，扩展到双目与大视角相机，实现了手机端的 AR 应用等其他功能[31~35]。


图 4　LSD-SLAM 的重构效果[20]

与 SVO 提取稀疏的特征点不同，单目 LSD-SLAM 系统利用图像中像素梯度显著的区域进行位姿跟踪和深度重构，可以恢复半稠密的三维场景地图，如图 4。整个算法主要分为三个线程，相机位姿跟踪、地图深度估计以及全局地图优化。在位姿跟踪线程中，当前帧 t 以上一帧图像的位姿作为初始估计，基于环境亮度一致性假设与距离最近的关键帧 k 进行比较，得到当前帧的相对位姿。

$$\xi_{tk}=\arg\min_{\xi_{tk}}\sum_{i\in\bar{R}}\left\|\frac{\delta I^2(u_i,\xi_{tk})}{\sigma^2_{\delta I(u_i,\xi_{tk})}}\right\|_\rho \quad (9)$$

其中：向量 ξ_{tk} 为变换矩阵 T_{tk} 的李代数表示[36]，$\delta I(u_i,\xi_{tk})$ 同式（6）。$\sigma^2_{\delta I}$ 为 $\delta I(u_i,\xi_{tk})$ 的方差，用于减小测量噪声对位姿跟踪的影响。

在 LSD-SLAM 的地图深度估计线程中，会根据当前帧 t 相对于关键帧 k 移动的距离 $dist(\xi_{t,k})$，判断是否将当前帧创建为新的关键帧。

$$dist(\xi_{t,k})=\xi_{t,k}^T W\xi_{t,k} \quad (10)$$

其中：W 为权重矩阵。若位移量超过设定的阈值，当前帧就会成为新的关键帧，得到从上一关键帧投影而来的半稠密像素点集。若位移量没有超过设定的阈值，当前帧就会从帧序列中选择一帧图像作为参考帧，进行极线搜索，得到当前帧中显著像素点的深度观测值 $d_t(u_i)$ 与方差 σ_t^2，再采用 EKF 更新关键帧中像素点的深度估计值与方差：

$$\begin{cases} d_k'(u_i)\leftarrow\dfrac{V_k(u_i)d_t(u_i)+\sigma_t^2 d_k(u_i)}{V_k(u_i)+\sigma_t^2} \\ V_k'(u_i)\leftarrow\dfrac{V_k(u_i)\sigma_t^2}{V_k(u_i)+\sigma_t^2} \end{cases} \quad (11)$$

因为单目相机不能计算像素点的绝对深度，会带来尺度漂移问题，所以 LSD-SLAM 对前两个线程创建的全局地图进行优化。对于保存的关键帧，LSD-SLAM 通过规定所有像素点逆深度的均值为 1 的办法控制全局地图的尺度；这样相邻关键帧之间就可以通过带有尺度变化的李代数表示其相对位姿关系[36]。LSD-SLAM 利用场景深度与跟踪精度的内在联系最小化误差，得到两个关键帧之间的相似变换 ζ_{ji}：

$$\zeta_{ji}=\arg\min_{\zeta_{j,i}}\sum_{i\in\bar{R}}\left\|\frac{\delta I^2(u_i,\zeta_{ji})}{\sigma^2_{\delta I(u_i,\zeta_{ji})}}+\frac{\delta d^2(u_i,\zeta_{ji})}{\sigma^2_{\delta d(u_i,\zeta_{ji})}}\right\|_\rho \quad (12)$$

其中：$\delta d(u_i,\zeta_{ji})$ 为两关键帧中相同像素点的深度差，$\sigma_{\delta d(u_i,\zeta_{ji})}$ 为该深度差的方差。最后，LSD-SLAM 依据关键帧之间的位姿关系和图像内容进行回环检测[37]，并采用捆集调整算法优化全局地图[15]。

2.3　DSO 算法分析

DSO 是 Engel 于 2016 年发布的一种单目稀疏直接法视觉里程计，在 2017 年扩展了双目功能[38]。与 SVO 一样，它不是完整的 SLAM 系统，没有回环检测模块。不同于特征点法需要

通过匹配特征点关联数据，DSO 将数据关联与位姿跟踪放在一个统一的非线性优化框架中求解。独特的相机光度标定模型[39]和滑动窗口优化[40, 41]的应用，使得 DSO 拥有良好的运算速度与跟踪精度，将直接法 V-SLAM 推向了一个新的高度。

由于直接法根据图像的灰度值跟踪相机的位姿，易受光照变化干扰。DSO 中提出了光度标定，对相机的晕影（图 5）、曝光时间、伽马响应[42]进行标定，补偿它们的影响。晕影是因为透镜之间的遮挡导致光强由图像中心向周围逐渐减小的现象，由一个权重映射 $V:\Omega\rightarrow[0,1]$ 标定；由于相机的曝光模式和快门打开方式不一，导致不同场景曝光时间不同，由曝光时间 t 表示；伽马响应非线性地映射相机的输入曝光量和输出灰度值之间的关系，由非线性响应函数 $G:\mathrm{R}\rightarrow[0,255]$ 标定。灰度映射模型为

$$I_i(\boldsymbol{u})=G(t_iV(\boldsymbol{u})B_i(\boldsymbol{u})) \tag{13}$$

其中：B_i 和 I_i 分别表示第 i 帧图像的辐射量和灰度值。DSO 算法的第一步就是通过方程式（13）进行灰度标定，利用标定之后的参数进行位姿跟踪。

与 LSD-SLAM 类似，DSO 的前端线程通过匹配新帧与关键帧来跟踪相机的位姿，然后通过一定的条件判断新帧是否会成为关键帧，并将新关键帧插入后端优化线程。DSO 的后端在一个滑动窗口内对维护的 5~7 个关键帧（图 6）进行优化。在各关键帧中提取深度收敛的像素点，投影至其他关键帧中，得到光度误差。

$$E_{uj}=\sum_{\boldsymbol{u}}w_{\boldsymbol{u}}\left\|(I_j[\boldsymbol{u}']-b_j)-\frac{t_je^{a_j}}{t_ie^{a_i}}(I_i[\boldsymbol{u}]-b_i)\right\|_\rho \tag{15}$$

$\boldsymbol{u}$、$\boldsymbol{u}'$ 为像素点在两帧中的图像坐标，t_i、t_j 为图像 I_i、I_j 的曝光时间。a_i、a_j、b_i、b_j 为光度仿射参数，在无法进行光度标定的情况下近似实现光度标定效果。

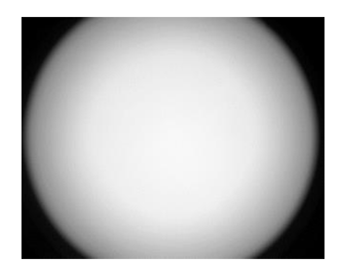


图 5　镜头晕影

最后滑动窗口通过优化光度误差之和

$$E_{photo}=\sum_{i\in F}\sum_{\boldsymbol{u}\in U_i}\sum_{j\in obs(\boldsymbol{u})}E_{uj} \tag{15}$$

得到优化的关键帧与像素点的位置，进而维护一个全局地图。式中，F 表示滑动窗口内的关键帧集合，U_i 表示关键帧 i 中提取的像素点集合，$obs(\boldsymbol{u})$ 表示滑动窗口内可以观察到像素点 $\boldsymbol{u}$ 的关键帧集合。

与传统的捆集调整对全局的关键帧和特征点进行优化不同，DSO 中的滑动窗口优化保存一定数量的关键帧，需要将多余的帧进行边缘化处理。滑动窗口优化中的边缘化通过更新信息矩阵[43]，将被删除帧的信息作为先验信息保存在信息矩阵中，控制了优化的计算量，又实现了良好的优化效果。

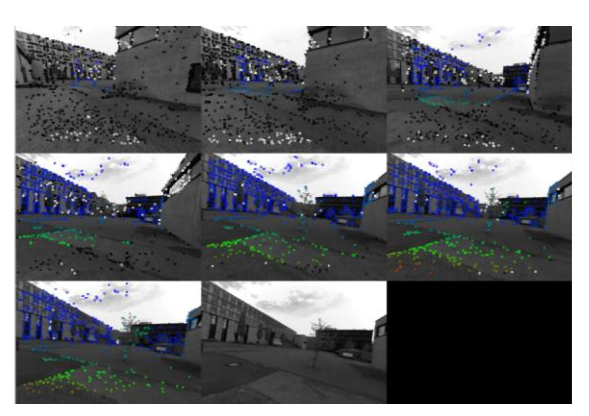


图 6　DSO 滑动窗口内的关键帧

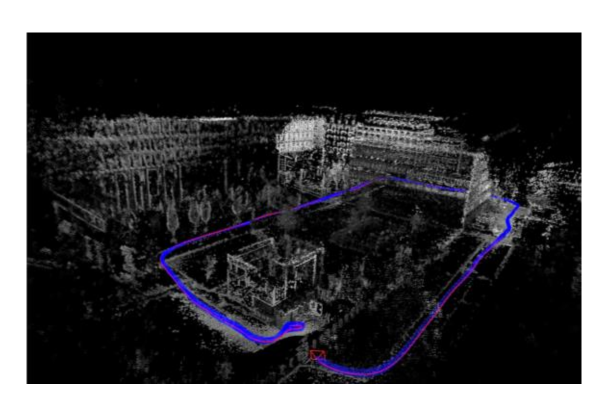


图 7　DSO 重建效果

2.4　分析比较

在单目实现的情况下，对上述各 SLAM 方法进行比较，结果如表 1 所示。

表 1　直接法 V-SLAM 系统分析比较

	SVO	LSD-SLAM	DSO
跟踪精度	★★	★	★★★
算法效率	★★★	★	★★
光照变化鲁棒性	★	★	★★★
快速运动鲁棒性	★★	★	★★
重定位能力	★	★★★	×
回环闭合能力	×	★★	×

1）跟踪精度

SVO 是一种半直接法视觉里程计，仅在前端的稀疏特征点匹配上使用了直接法，后端中仍使用重投影误差维护构建的地图，因此跟踪精度会优于以直接像素匹配进行位姿跟踪的 LSD-SLAM。LSD-SLAM 通过匹配像素进行位姿跟踪，易受环境光照变化和相机曝光等因素的影响，跟踪精度较低；但由于它是完整的 SLAM 系统可以进行回环检测[37]，通过全局优化在一定程度上弥补精度不足的缺点。DSO 创新性地提出了光度标定方

法，很大程度上解决了相机曝光等因素对直接像素匹配的影响，滑动窗口优化可以有效地消除误差累计，定位精度较高。在只使用视觉里程计模块的情况下，根据文献[26]在 EuRoc 数据集[44]上运行的结果,DSO 的跟踪精度高于 SVO，LSD-SLAM 的跟踪精度最不理想。

2）算法效率

SVO 通过直接图像匹配巧妙地避免了耗时的描述子计算与匹配，又因为提取的特征点稀疏，运算效率极高，在 CPU（i7 处理器/3 GHz）上可以达到了 300 fps 左右的处理速度，不需占用大量计算资源。LSD-SLAM 对图像中像素梯度较为明显的区域提取像素点，是一种半稠密直接法，需要求解的信息矩阵较为稠密，在一定程度上影响了算法的效率。另外，LSD-SLAM 需要对关键帧进行回环检测，所以算法效率较低，在 CPU（i7 处理器/3 GHz）上的处理速度为 20-30 fps 左右。DSO 是稀疏直接法，良好的图像匹配能力使得它可以通过降低图像分辨率达到较高的运算速度；滑动窗口优化的使用控制了后端处理的关键帧数，计算量较低。DSO 的跟踪精度与算法效率之间存在权衡，通常图像分辨率越高跟踪精度越高，但算法效率越低。在同等跟踪精度下，DSO 的算法效率优于 LSD-SLAM。

3）光照变化鲁棒性

直接法通过比较图像信息来实现位姿跟踪，容易受光照变化的影响。SVO 对图像中像素梯度较大的稀疏点进行计算，容易受环境光照不一致变化和相机曝光时间不同等因素的影响。LSD-SLAM 对光照变化的鲁棒性也受制于上述因素。光度标定和动态估计的光度参数使得 DSO 对相机成像引起的图像明暗变化会有更鲁棒的效果，但对于环境光照不一致变化的作用有限。

4）快速运动鲁棒性

能否在相机快速运动的情况下跟踪相机的位姿，体现了 SLAM 算法对相机快速运动的鲁棒性。不同于特征点法中特征点描述子可以在相机大运动的情况下进行匹配的特点，直接法假设相机平缓运动，在位姿跟踪中需要有一个良好的相机位姿作为初始值进行图像匹配。因此直接法的快速运动鲁棒性都不是很好。LSD-SLAM 在跟踪时要处理半稠密的像素点，对相机运动速度最为敏感，快速运动鲁棒性不好。SVO 的跟踪效率很高，相机位姿预测和基于图像金字塔模型的匹配在一定程度上弥补了快速运动鲁棒性的不足。DSO 可以通过调节图像分辨率实现快速运动跟踪，鲁棒性优于 LSD-SLAM。

5）重定位/回环闭合能力 系统在实际运行中会出现位姿跟踪丢失的情况，重定位模块可以恢复丢失的位姿。DSO 中没有重定位模块。SVO 通过维护关键帧组成的局部地图，在跟踪丢失的情况下将当前帧与最近的关键帧进行匹配恢复初始位姿，再把局部地图中的特征点投影到当前帧进行特征匹配和位姿优化来实现重定位。该方法鲁棒性不足，当前帧与关键帧的位姿变化较大时，就不容易重定位成功。LSD-SLAM 在位姿跟踪丢失的情况下，使用特征点描述子和 FAB-MAP 检索方法[37]来实现重定位。因为特征点描述子对视角变化具有不变性，所以 LSD-SLAM 重定位的鲁棒性较好。DSO 和 SVO 都没有回环检测模块。LSD-SLAM 通过特征点和 FAB-MAP 方法检测回路，在回环发生时通过位姿图优化全局位姿，回路闭合能力较强。

3　直接法 V-SLAM 的优缺点与发展方向

直接法基于特殊的环境亮度不变性假设建立了数据之间的关联，通过图像灰度值构建了目标函数中的残差，因此它的优缺点也十分明显。

3.1　直接法的优点

直接法无须进行耗时的 ORB[45]、SIFT[46]、SURF[47]等特征点和描述子的提取与匹配过程，因此运算效率较高。在相同的计算资源条件下，相比于特征点法几十帧每秒的运算能力，直接法普遍可以达到上百帧每秒的处理速度。并且与特征点法只提取图像中几百个像素点作为特征点不同，直接法使用图像中几十万个像素点的信息，对存在像素梯度的区域进行匹配，在特征缺失、纹理重复的环境中仍可以准确跟踪。大量图像信息的使用，也使得直接法可以恢复稠密或半稠密的场景地图。另外，无须特征匹配，直接法以更加整体和鲁棒的方式进行数据关联，避免了特征点误匹配给 SLAM 系统带来的致命影响，可以提供更高的精度和系统稳定性。

3.2　直接法的缺点

直接法的非线性优化基于图像像素梯度，对代价函数使用梯度下降求解最优值。因为图像像素的非凸性，优化过程容易陷入局部极小值，所以直接法需要一个不错的初始位姿估计值和较高的图像质量。因此，在相机运动速度较快和拍摄帧率不高的情况下，位姿跟踪容易丢失。另外，环境光度一致性假设是一种较强的假设，对环境的光照要求较高。相机曝光时间的变化、快门的打开方式、相机对曝光的调节和环境光照条件的变化都会使得跟踪算法失效。因此需要使用特殊功能的相机（如基于事件的相机 event-camera[48-49]）、全局曝光的镜头（global shutter）和光度标定方法。最后，直接法中无法使用特征点，不能如特征点法一样对存储的特征点和描述子进行匹配实现重定位和回环检测，需要创新直接法的重定位、回环检测模块。

3.3　直接法的发展方向

直接法的提出为解决 V-SLAM 问题提供了一条新的思路，但还未达到成熟的水平，仍存在许多问题可以开展进一步的研究。

3.3.1 多传感器融合

直接法 V-SLAM 使用相机作为传感器，易受光照影响，在镜头遮挡、相机抖动严重的情况下位姿跟踪容易丢失。传统的定位方法也存在许多不足。使用 IMU 惯性器件进行长时间定位，累计误差严重；民用 GPS 的精度较差，在室内和室外遮挡严重的情况下难以获得位置信号。基于单一传感器的定位都存在各自的局限性，多传感器的融合可以提高系统的精度与鲁棒性。相机与 IMU 惯性器件融合，可以将相机采集的丰富图像信

息和 IMU 惯性器件的短时精确测量数据进行耦合[50-52]，实现良好的 SLAM 效果。直接法基于像素梯度进行位姿跟踪，需要良好的相机位姿初始值进行优化，IMU 惯性器件可以很好地解决这一问题。另外，精确的位姿跟踪也可以用于消除 IMU 惯性器件的 bias 漂移。但与其他类型传感器的融合会增加直接法 V-SLAM 优化线程中的参数量，增加信息矩阵的稠密度，降低系统的实时性。为了解决优化参数过多的问题，可以使用滑动窗口优化的方法，将收敛的参数边缘化，作为先验信息传递给后续待优化的参数。目前，直接法 V-SLAM 与 IMU 惯性器件融合的工作主要有文献[26,35]。

3.3.2 回环检测/重定位模块

直接法V-SLAM直接使用图像的像素信息进行位姿跟踪与地图建立，缺少了图像特征提取过程，难以如特征点法一样，通过匹配图像特征来进行回环检测和重定位。为了解决直接法难以实现地图重用的问题，目前主要有两种研究思路。一种如 LSD-SLAM 中提取特征点进行回环检测和重定位一样，可以将特征点法与直接法的优势结合到一起。利用特征点法提取的特征实现地图重用功能，结合直接法跟踪效率高、系统鲁棒和使用图像信息丰富的优点，创建一个融合的系统。另一种思路就是利用深度学习方法，匹配关键帧图像实现地图重用。深度学习应用神经网络提取图像中识别率更高的深层次特征[53]。深度学习特征的使用可以提升 SLAM 系统回环检测和重定位的准确率。Kendall 在 2015 年提出了 PoseNet，实现了相机位姿的实时重定位[54]。PoseNet 卷积神经网络利用带有重构三维模型和相机位姿的图像数据对神经网络进行训练，采用端到端的方法对相机的位姿进行重定位。类似的工作还有文献[55~57]。

3.3.3 直接法 V-SLAM 与深度学习融合

深度学习是近年来的研究热点。作为一种端到端的方法，深度学习可以替代 SLAM 系统中的某一个模块，或者采用非传统的框架直接解决机器人导航问题[58,59]。在深度学习与直接法 V-SLAM 结合方面，Zhou 提出的 SfM-Learner[60]基于光度投影一致性假设，使用 CNN 来估计每一帧图像的深度信息并跟踪相机位姿，替代了传统 SLAM 系统前端中的部分模块。SfM-Net[61]在 SfM-Learner 的基础上计算了光流和三维点云。类似的工作还有[62]。CNN-SLAM[63]使用直接法跟踪相机位姿，结合 CNN 来估计场景深度并对图像进行语义分割，得到结合几何与语义信息的场景地图。另外文献[64,65]在特征点提取、语义分割等方面也做了一定的工作。SLAM 问题是一个具有严格闭环数学表达的几何问题，而深度学习更擅长在回环检测、图像分割、图像语义等 SLAM 更高级的应用上开展工作，两者的结合可以推动 SLAM 技术的发展。

4 结束语

近年来，SLAM 技术越来越多地被应用在虚拟现实终端、自主航行无人机以及无人驾驶汽车等硬件平台。各式硬件传感器以及高性能图像处理单元的发展与普及，推动着 SLAM 技术朝着高精度、强鲁棒、多传感器融合的方向发展。

直接法 V-SLAM 技术的提出，为解决 SLAM 问题提供了一条新的思路，弥补了一些特征点法 V-SLAM 技术的不足。但直接法仍存在许多缺点，受限于诸如环境光照条件、相机运动条件、地图重用困难等情况的限制。这就对直接法 V-SLAM 算法提出了更高的鲁棒性和功能拓展的要求。如何进一步提高直接法 V-SLAM 对环境光照和相机运动的鲁棒性，实现地图重用功能将是一个很有价值的研究方向。此外，深度学习与 SLAM 技术的结合在一定程度上改善了传统 SLAM 算法的局限性，为 SLAM 更高级的应用提供了思路。将深度学习应用到直接法 V-SLAM 系统中将是一个充满意义又富有挑战的研究方向。

参考文献：

[1] Durrant-Whyte H, Bailey T. Simultaneous localization and mapping: Part I [J]. IEEE Robotics & Automation Magazine, 2006, 13 (2): 99–108.

[2] Bailey T, Durrant-Whyte H. Simultaneous localization and mapping (SLAM): Part II [J]. IEEE Robotics & Automation Magazine, 2006, 13 (3): 108-117.

[3] Fuentes-Pacheco J, Ruiz-Ascencio J, Rendón-Mancha J M. Visual simultaneous localization and mapping: a survey [J]. Artificial Intelligence Review, 2015, 43 (1): 55-81.

[4] Bonin-Font F, Ortiz A, Oliver G. Visual navigation for mobile robots: a survey [J]. Journal of Intelligent and Robotic Systems, 2008, 53 (3): 263-296.

[5] Faessler M, Fontana F, Forster C, *et al.* Autonomous, vision-based flight and live dense 3D mapping with a quadrotor micro aerial vehicle [J]. Journal of Field Robotics, 2016, 33 (4): 431-450.

[6] Liu H M, Zhang G F, Bao H J. A survey of monocular simultaneous localization and mapping [J]. Journal of Computer-Aided Design and Computer Graphics, 2016, 28 (6): 855–868.

[7] Smith R C, Cheeseman P. On the representation and estimation of spatial uncertainty [J]. International Journal of Robotics Research, 1986, 5 (4): 56-68.

[8] Smith R, Self M, Cheeseman P. Estimating uncertain spatial relationships in robotics [J]. Machine Intelligence & Pattern Recognition, 1990, 4 (5): 435-461.

[9] Thrun S, Burgard W, Fox D. Probabilistic Robotics [M]. Cambridge: MIT Press, 2005.

[10] Huang G Q, Mourikis A I, Roumeliotis S I. Analysis and improvement of the consistency of extended Kalman filter based slam [C]// Proc of IEEE International Conference on Robotics and Automation. 2008: 473-479.

[11] Martines-Cantin R, Castellanos J A. Unscented SLAM for large-scale outdoor environments [C]// Proc of IEEE//RSJ International Conference On Intelligent Robots and Systems. 2005: 401-422.

[12] Montemerlo M. FastSLAM: A Factored solution to the simultaneous localization and mapping problem with unkown data association [D].

Pittsburgh, PA: Carnegie Mellon University, 2003.

[13] Davison A J, Reid I D, Molton N D, *et al.* Monoslam: real-time single camera slam [J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2007, 29 (6): 1052-1067.

[14] Hartley R, Zisserman A. Multiple view geometry in computer vision [M]. Cambridge: Cambridge University Press, 2004.

[15] Triggs B, McLauchlan P F, Hartley R I, *et al.* Bundle adjustment: a modern synthesis [C]// Proc of International Workshop on Vision Algorithms. Berlin: Springer, 1999: 298–372.

[16] Klein G, Murry D. Parallel tracking and mapping for small AR workspaces [C]// Proc of IEEE and ACM International Symposium on Mixed and Augmented Reality. Los Alamitos: IEEE Computer Society Press, 2007: 225-234.

[17] Klein G, Murry D. Parallel tracking and mapping on a camera phone [C]// Proc of IEEE and ACM International Symposium on Mixed and Augmented Reality. Los Alamitos: IEEE Computer Society Press, 2009: 83-86.

[18] Mur-Artal R, Montiel J M M, Tardós J D. ORB-SLAM: a versatile and accurate monocular SLAM system [J]. IEEE Trans on Robotics, 2015, 31 (5): 1147-1163.

[19] Mur-Artal R, Tardós J D. ORB-SLAM2: an open-source SLAM system for monocular, stereo, and RGB-D cameras [J]. IEEE Trans on Robotics, 2016, 33 (5): 1255-1262.

[20] Engel J, Schöps T, Cremers D. LSD-SLAM: large-scale direct monocular SLAM [C]// Proc of Computer Vision-ECCV. Heidelberg: Springer, 2014: 834-849.

[21] 高翔，张涛，等. 视觉SLAM十四讲 [M]. 北京：电子工业出版社, 2017. (Gao Xiang, Zhang Tao, *et al.* The fourteen lectures on visual SLAM [M]. Beijing: Publishing House of Electronics Industry, 2017.)

[22] Forster C, Pizzoli M, Scaramuzza D. SVO: fast semi-direct monocular visual odometry [C]// Proc of IEEE International Conference on Robotics and Automation. 2014: 15-22.

[23] Engel J, Koltun V, Cremers D. Direct sparse odometry [J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2016, pp (99): 1-14.

[24] Newcombe R A, Lovegrove S J, Davison A J. DTAM: dense tracking and mapping in real-time [C]// Proc of International Conference on Computer Vision. 2011: 2320-2327.

[25] Kerl C, Sturm J, Cremers D. Robust odometry for RGB-D cameras [C]// Proc of IEEE International Conference on Robotics and Automationa. 2013: 3748-3754.

[26] Forster C, Zhang Z C, Michael G, *et al.* SVO: semidirect visual odometry for monocular and multicamera systems [J]. IEEE Trans on Robotics, 2017, 33 (2): 249-265.

[27] Civera J, Davison A J, Montiel J M. Inverse depth parametrization for monocular SLAM [J]. IEEE Trans on Robotics, 2008, 24 (5): 932-945.

[28] Vogiatzis G, Hernandez C. Video-based, real-time multi-view stereo [J]. Image & Vision Computing, 2011, 29 (7): 434-441.

[29] Pizzoli M, Forster C, Scaramuzza D. REMODE: probabilistic, monocular dense reconstruction in real time [C]// Proc of IEEE International Conference on Robotics and Automation. 2014: 2609-2616.

[30] Engel J, Cremers D. Semi-dense visual odometry for a monocular camera [C]// Proc of IEEE International Conference on Computer Vision. 2013: 1449-1456.

[31] Schops T, Enge J, Cremers D. Semi-dense visual odometry for AR on a smartphone [C]// Proc of IEEE International Symposium on Mixed and Augmented Reality. Washington DC: IEEE Computer Society, 2014: 145-150.

[32] Engel J, Stückler J, Cremers D. Large-scale direct SLAM with stereo cameras [C]// Proc of IEEE//RSJ International Conference on Intelligent Robots and Systems. 2015: 1935-1942.

[33] Caruso D, Engel J, Cremers D. Large-scale direct SLAM for omnidirectional cameras [C]// Proc of IEEE//RSJ International Conference on Intelligent Robots and Systems. 2015: 141-148.

[34] Usenko V, Engel J, Stückler J, *et al.* Reconstructing street-scenes in real-time from a driving car [C]// Proc of International Conference on 3D Vision. Washington DC: IEEE Computer Society, 2015: 607-614.

[35] Usenko V, Engel J, Stückler J, *et al.* Direct Visual-Inertial Odometry with Stereo Cameras [C]// Proc of IEEE International Conference on Robotics and Automation. 2016: 1885-1892.

[36] Barfoot T D. State estimation for robotics: a matrix Lie group approach [M]. Cambridge: Cambridge University Press, 2017.

[37] Glover A, Maddern W, Warren M, *et al.* OpenFABMAP: an open source toolbox for appearance-based loop closure detection [C]// Proc of IEEE International Conference on Robotics and Automation. 2012: 4730-4735.

[38] Wang R, Schwörer M, Cremers D. Stereo DSO: large-scale direct sparse visual odometry with stereo cameras [C]// Proc of IEEE International Conference on Computer Vision. 2017: 3923-3931.

[39] Engel J, Usenko V, Cremers D. A photometrically calibrated benchmark for monocular visual odometry [EB/OL]. (2016-10-08) [2017-10-24]. http://arxiv. org/abs/1607. 02555.

[40] Stefan L, Simon L, Michael B, *et al.* Keyframe-based visual-inertial odometry using nonlinear optimization [J]. International Journal of Robotics Research, 2015, 34 (3): 314-334.

[41] Sibley G, Matthies L, Sukhatme G. Sliding window filter with application to planetary landing [J]. Journal of Field Robotics, 2010, 27 (5): 587–608.

[42] Debevec P E, Malik J. Recovering high dynamic range radiance maps from photographs [C]// Proc of Conference on Computer Graphics and Interactive Techniques. 1997: 369-378.

[43] Eckenhoff K, Paull L, Huang G. Decoupled, consistent node removal and edge sparsification for graph-based SLAM [C]// Proc of IEEE//RSJ International Conference on Intelligent Robots and Systems. 2016: 3275-3282.

[44] Burri M, Nikolic J, Gohl P, *et al.* The EuRoC micro aerial vehicle datasets [J]. International Journal of Robotics Research, 2016, 35 (10): 1157-1163.

[45] Rublee E, Rabaud V, Konolige K, *et al.* ORB: an efficient alternative to SIFT or SURF [C]// Proc of IEEE International Conference on Computer Vision. 2011: 2564-2571.

[46] Lowe D G. Distinctive Image Features from Scale-Invariant Keypoints [J]. International Journal of Computer Vision, 2004, 60 (2): 91-110.

[47] Bay H, Tuytelaars T, Gool L V. SURF: speeded up robust features [J]. Computer Vision & Image Understanding, 2006, 110 (3): 404-417.

[48] Kim H, Handa A, Benosman R, *et al.* Simultaneous mosaicing and tracking with an event camera [C]// Proc of the British Machine Vision Conference. 2014: 1-12.

[49] Rebecq H, Horstschaefer T, Gallego G, *et al.* EVO: a geometric approach to event-based 6-DOF parallel tracking and mapping in real time [J]. IEEE Robotics & Automation Letters, 2017, 2 (2): 593-600.

[50] Weiss S M. Vision based navigation for micro helicopters [D]. Zurich, Switzerland: ETH Zurich, 2012.

[51] Mourikis A I, Roumeliotis S I. A Multi-State Constraint Kalman Filter for Vision-aided Inertial Navigation [C]// Proc of IEEE International Conference on Robotics and Automation. 2007: 3565-3572.

[52] Li M Y, Mourikis A I, Anastasios I. High-precision, consistent EKF-based visual–inertial odometry [J]. International Journal of Robotics Research, 2013, 32 (6): 690-711.

[53] Girshick R, Donahue J, Darrel T, *et al.* Rich feature hierarchies for accurate object detection and semanic segmentation [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2014: 580-587.

[54] Kendall A, Grimes M, Cipolla R. PoseNet: A convolutional network for real-time 6-DOF camera relocalization [C]// Proc of IEEE International Conference on Computer Vision. 2015: 2938-2946.

[55] Wu J, Ma L, Hu X L. Delving deeper into convolutional neural networks for camera relocalization [C]// Proc of IEEE International Conference on Robotics and Automation. 2017.

[56] Chen Z, Jacobson A, Sunderhauf N, *et al.* Deep learning features at scale for visual place recognition [C]// Proc of IEEE International Conference on Robotics and Automation. 2017: 3223-3230.

[57] Gao X, Zhang T. Unsupervised learning to detect loops using deep neural networks for visual SLAM system [J]. Autonomous Robots, 2017, 41 (1): 1-18.

[58] Zhu Y, Mottaghi R, Kolve E, *et al.* Target-driven visual navigation in indoor scenes using deep reinforcement learning [EB/OL]. (2016-09-16) [2017-10-24]. http://arxiv. org/abs/1609. 05143.

[59] Gupta S, Davidson J, Levine S, *et al.* Cognitive mapping and planning for visual navigation [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. Washington DC: IEEE Computer Society, 2017: 7272-7281.

[60] Zhou T, Brown M, Snavely N, *et al.* Unsupervised learning of depth and ego-motion from video [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. Washington DC: IEEE Computer Society, 2017: 6612-6619.

[61] Vijayanarasimhan S, Ricco S, Schmid C, *et al.* SfM-Net: learning of structure and motion from video [EB/OL]. (2017-04-25) [2017-10-24]. http://arxiv. org/abs/1704. 07804.

[62] Ummenhofer B, Zhou H, Uhrig J, *et al.* DeMoN: depth and motion network for learning monocular stereo [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. 2016: 5622-5631.

[63] Tateno K, Tombari F, Laina I, *et al.* CNN-SLAM: Real-time dense monocular SLAM with learned depth prediction [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. Washington DC: IEEE Computer Society, 2017: 6565-6574.

[64] Yi K M, Trulls E, Lepetit V, *et al.* LIFT: learned invariant feature transform [C]// Proc of European Conference on Computer Vision. Amsterdam: Springer, 2016: 467-483.

[65] Li X, Belaroussi R. Semi-dense 3D semantic mapping from monocular SLAM [EB/OL]. (2016-11-13) [2017-10-24]. http://arxiv. org/abs/1611. 04144.